IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

MAITE RUSCIEL GARCIA
ENERO,2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. We want to predict if the Falcon 9 first stage will land successfully. First of all, we collected the data with a request to the SpaceX API and web scrapping from a Wikipedia page. Then we used some basic data wrangling and formatting. With SQL and visualization techniques, the existing relationships between the variables were analyzed, machine learning classification models were applied to select the one that best fits, which turned out to be the decision tree. Here we show the work done and the results.

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- We want to predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

- Data collection methodology:

  For this project the data was collected in two ways, one part with a request call to the SpaceX API, and the other part from the Wikipedia page about Falcon 9 with web scrapping.

- Data wrangling:

  The data was processed to find the missing values, identify the categorical and numerical variables and obtained the outcome variable

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models and select the best fits.

Data Collection

For this project the data was collected in two ways:

- One part with a request call to the SpaceX API
- The other part from the Wikipedia page about Falcon 9 with web scrapping.

Data Collection – SpaceX API

- The chart represent the data collection from the SpaceX API:

GitHub URL of SpaceX API calls notebook:

https://github.com/mrusciel/spacey-data-science/blob/1b7ff021e188d895ea6fb9696e1223817798bfa6/jupyter-labs-spacex-data-collection-api.ipynb

Get Request
- Get request to the API url
- The json is saved in the response

Convert to Data Frame
- Tranform the json to data frame with pandas
- Select the specific columns we need.

Transform
- From those columns get the new data columns
- Create a dictionary with the new data
- Generate a new data frame with the dictionary.

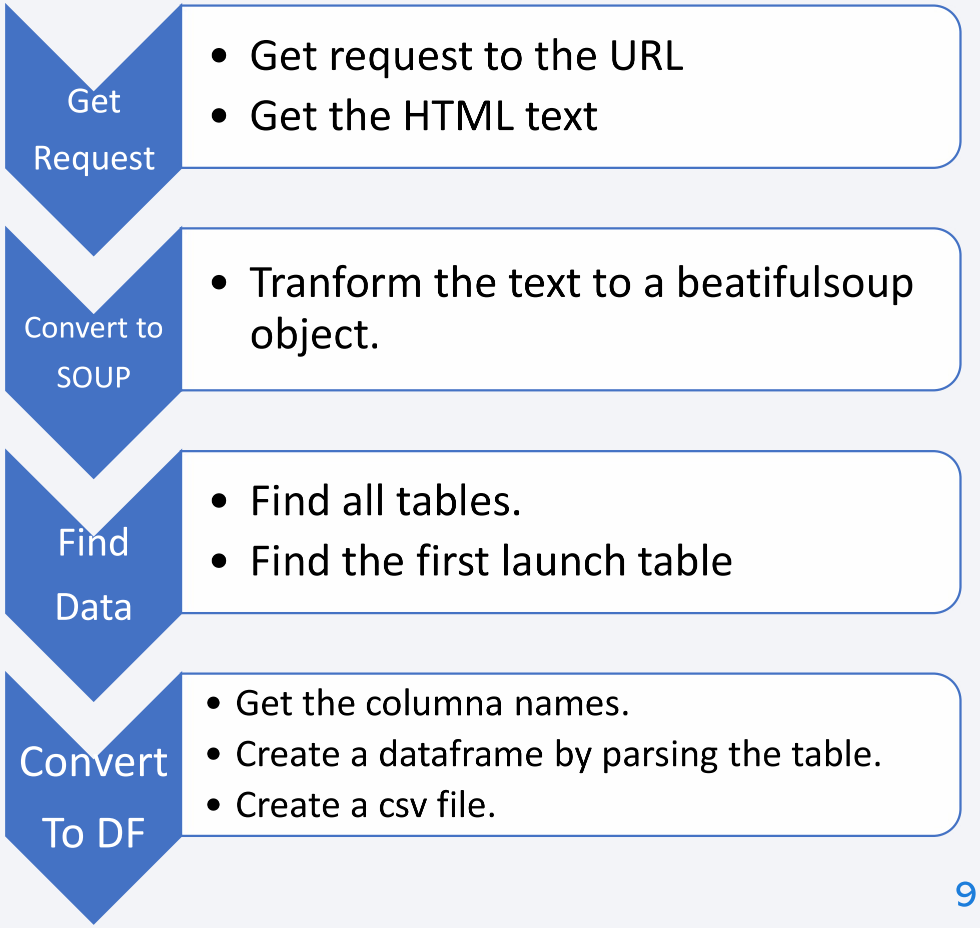
Filter
- Get only the Falcon 9 launch
- Reset the numbers of fligths

Data Collection - Scraping

- The chart represent the data collection from the Wikipedia page: SpaceX :

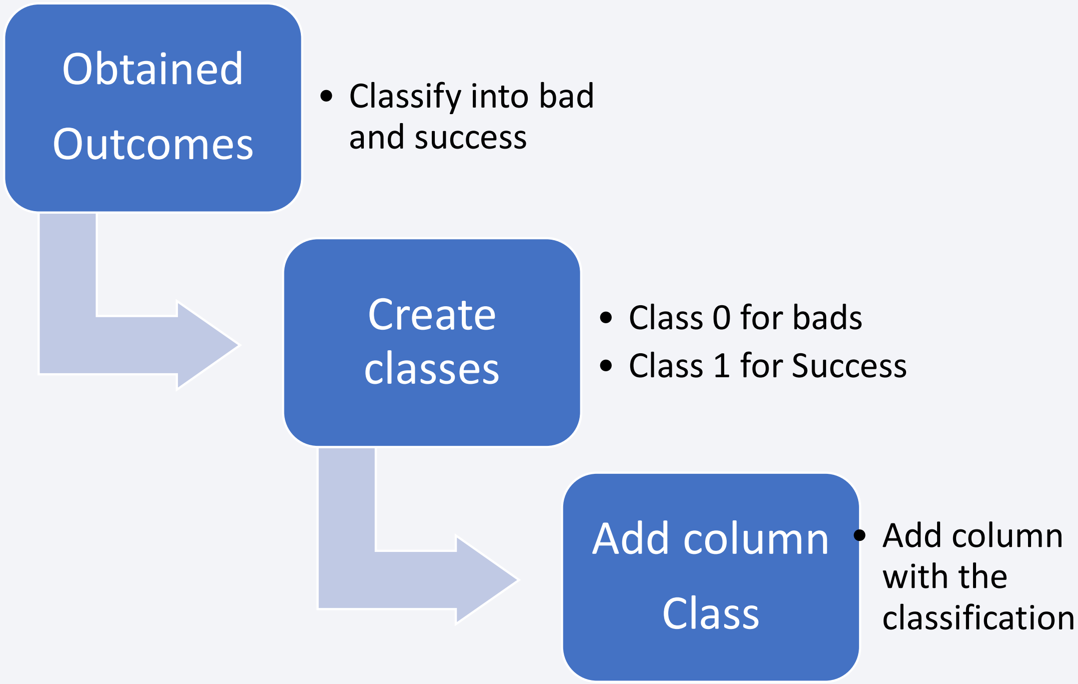
GitHub URL of the web scraping notebook:

https://github.com/mrusciel/spacey-data-science/blob/1b7ff021e188d895ea6fb9696e1223817798bfa6/jupyter-labs-webscraping.ipynb



Data Wrangling

- The data was processed to find the missing values, identify de categorical and numerical variables and obtained the outcome variable as it's shown in the following chart:



GitHub URL data wrangling related notebooks:

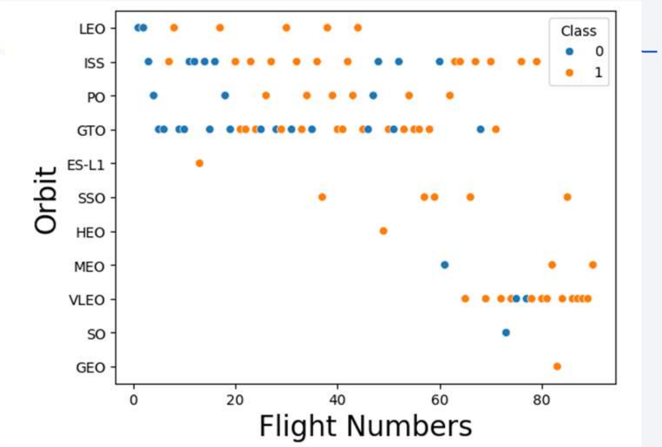
https://github.com/mrusciel/spacey-data-science/blob/1b7ff021e188d895ea6fb9696e1223817798bfa6/labs-jupyter-spacex-Data%20wrangling.ipynb

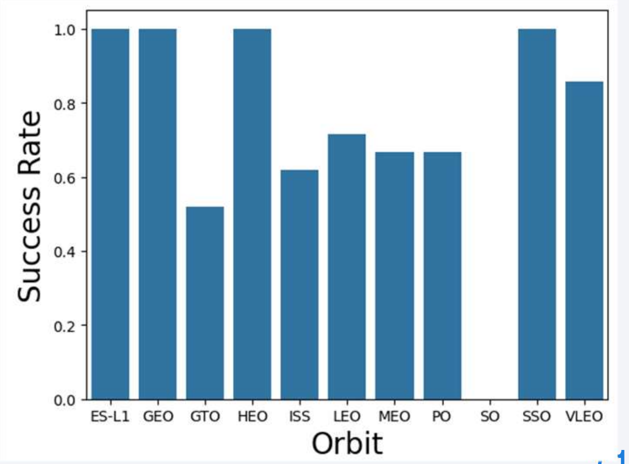
EDA with Data Visualization

- The data was analysed with some scatter plots, bar plots, and line plots to find the relationship between the launch site, the orbit, payload mass and the Success Rate of the launches. Here is some examples.

GitHub URL of EDA with data visualization notebook:

https://github.com/mrusciel/spacey-data-science/blob/e1bbc1c63b9ab6992ee82e4bbe7b41f7b1101408/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb



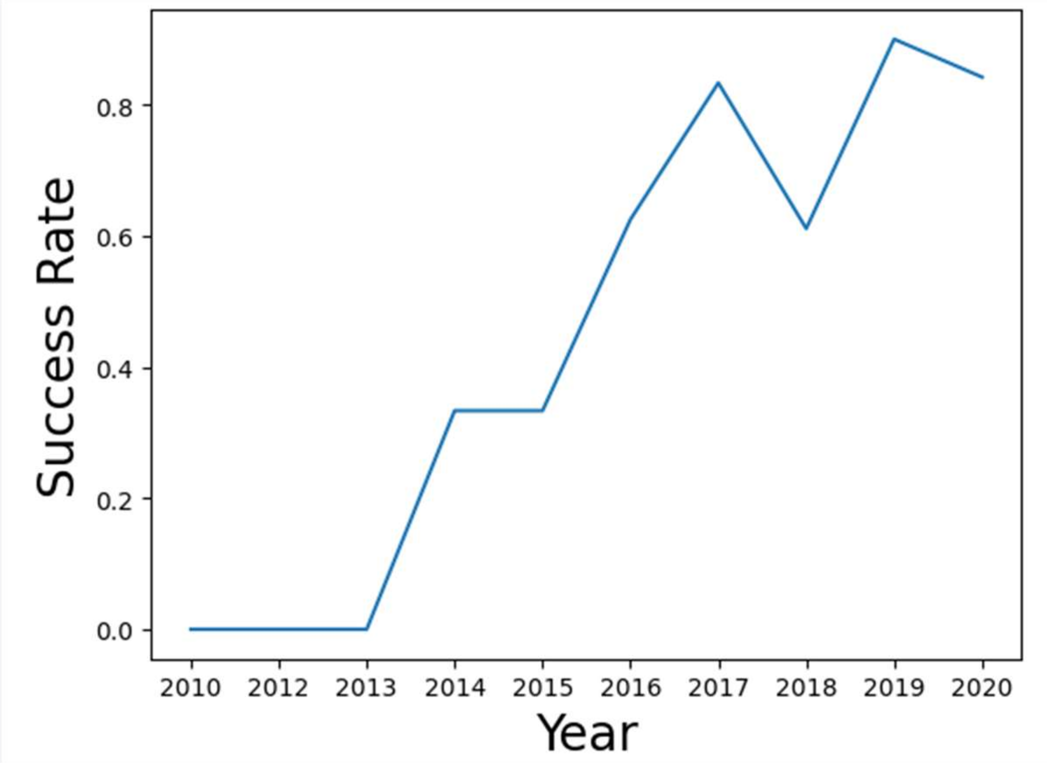


EDA with Data Visualization

- In this example we have the success rate by year, and we can see that in 2013 the rise began.

GitHub URL of EDA with data visualization notebook:

https://github.com/mrusciel/spacey-data-science/blob/e1bbc1c63b9ab6992ee82e4bbe7b41f7b1101408/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb



EDA with SQL

SQL was used to obtained:

- The names of the unique launch sites in the space mission
- The total payload mass carried by boosters launched by NASA (CRS)
- The average payload mass carried by booster version F9 v1.1
- The date when the first successful landing outcome in ground pad was achieved.
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- The total number of successful and failure mission outcomes.
- The names of the booster versions which have carried the maximum payload mass.
- The failure landing outcomes in drone ship in year 2015.
- The count of landing outcomes between the date 2010-06-04 and 2017-03-20

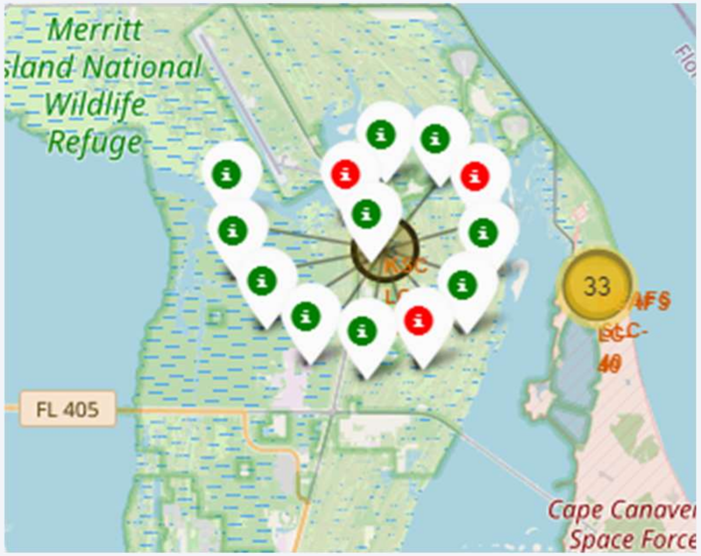
GitHub URL of my EDA with SQL notebook:

https://github.com/mrusciel/spacey-data-science/blob/471413dfbf278822a59ea378b83e8e5c67558880/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- To make the map we used the Folium library, circles and marks were positioned at the locations of the launch stations, and a marker cluster was used since some of them are very close. For each launch site we represented the success landings and failed landings.
- GitHub URL of my Folium map:

https://github.com/mrusciel/spacey-data-science/blob/e1bbc1c63b9ab6992ee82e4bbe7b41f7b1101408/lab_jupyter_launch_site_location.jupyterlite.ipynb

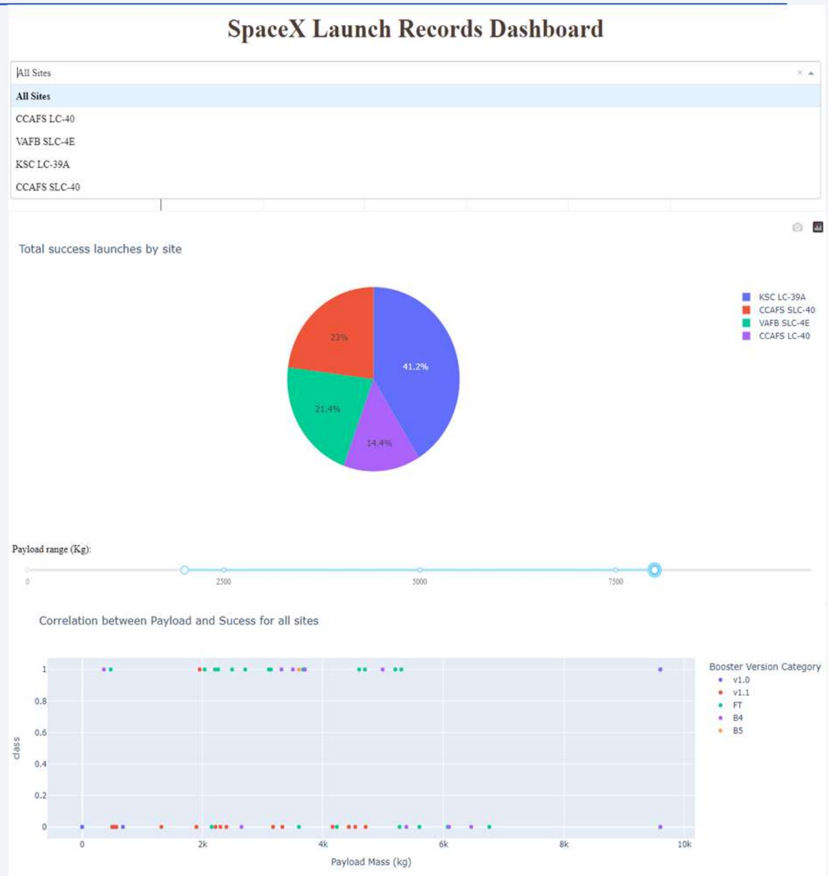


Build a Dashboard with Plotly Dash

- A dashboard was obtained that allows showing:
  - a pie chart with the range of successes by launch site
  - a scatter plot showing the relationship between launch site and payload mass
- Two interactions were inserted to select the launch site and the payload mass range to obtain independent graphs for each case.

GitHub URL Plotly Dash:

https://github.com/mrusciel/spacey-data-science/blob/1b7ff021e188d895ea6fb9696e1223817798bfa6/spacex_dash_app.py



Predictive Analysis (Classification)

Preparing datasets

- Read data frame from csv file.
- Split into dependent variables (X) and independent variables (Y).
- Split into train and test sets.

Modeling

- Apply each classification model.
- Get accuracy for each one.

Select the proper model

- Compare the accuracies and select the best model to predict.

- GitHub URL of predictive analysis:

https://github.com/mrusciel/spacey-data-science/blob/28be3ea9994fb9662094920e364cb03d5fc7fa4d/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

Next:

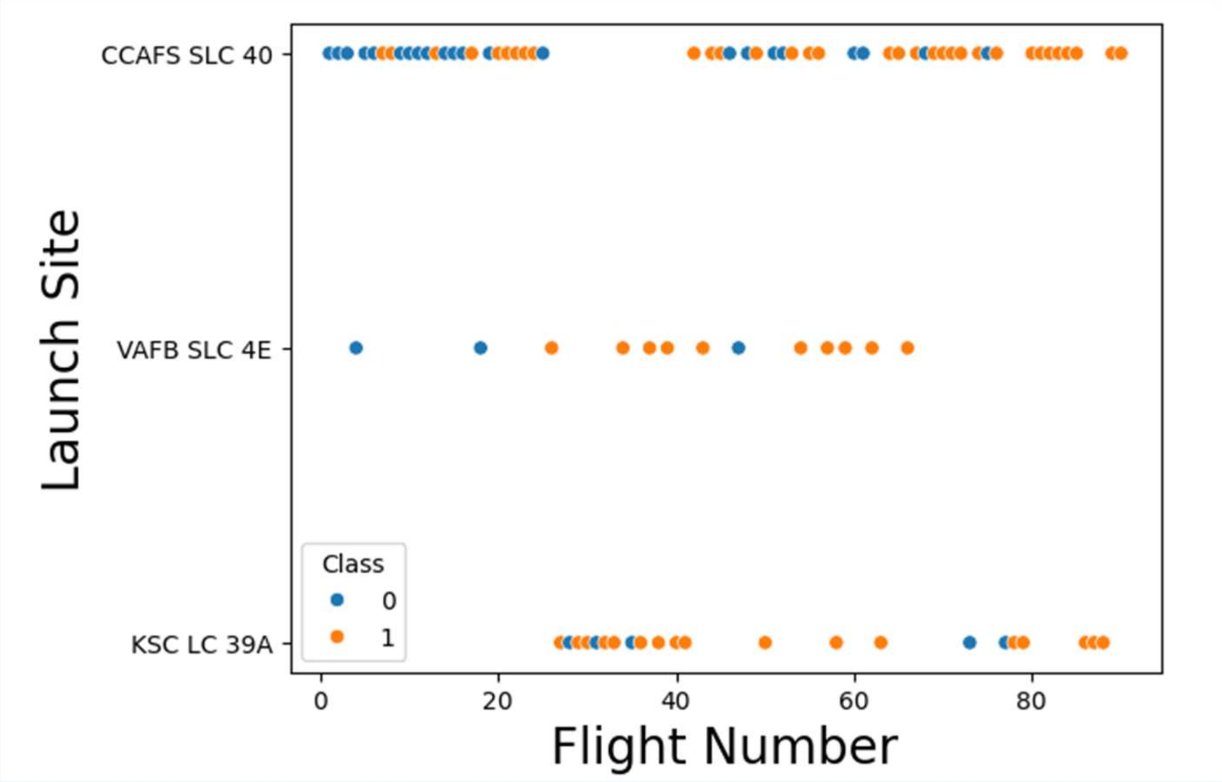
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results.

Section 2

Insights drawn from EDA

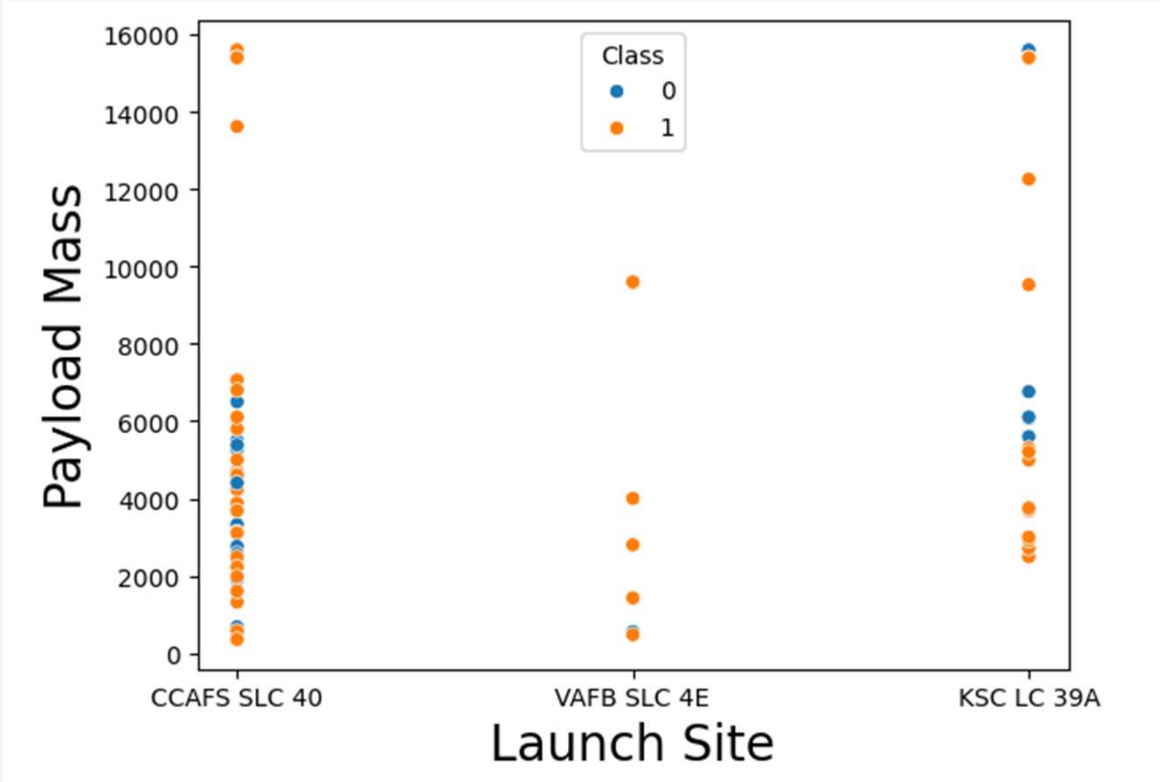
Flight Number vs. Launch Site

- Here we have a scatter plot for Flight Numbers vs Launch Sites.
- In orange the success landings and blue the failed landings for each site.
- As we can see in the site CCAFS SLC 40 the success rate improve in the last ones launches.



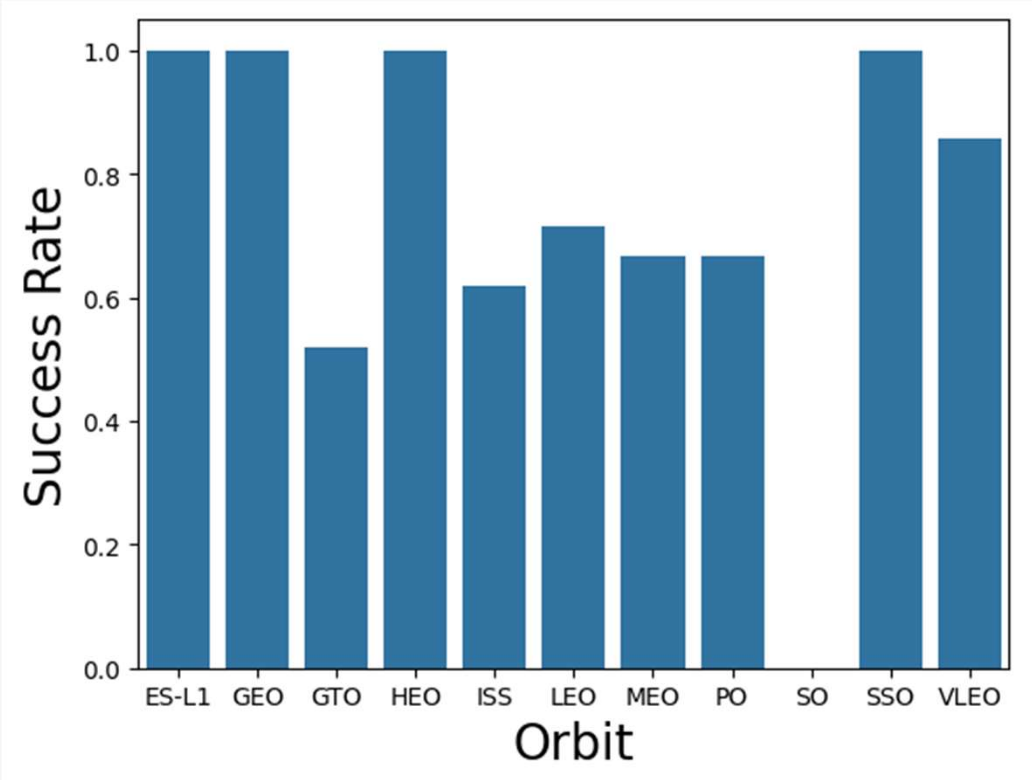
Payload vs. Launch Site

- Here we have a scatter plot for Launch Site vs Payload Mass.
- In orange color the success landings and blue the failed landings for each site.



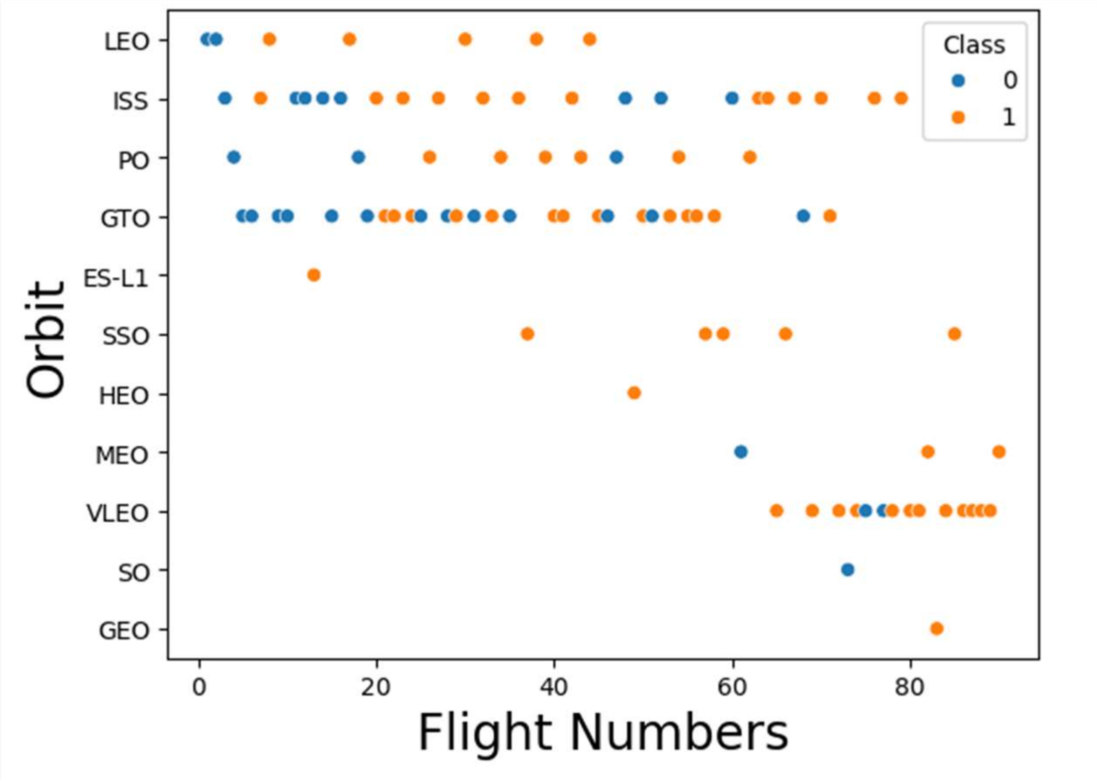
Success Rate vs. Orbit Type

- Here we have a bar plot for Orbit vs Success Rate.
- We can see there are four orbit with full success rate: ES-L1, GEO, HEO, SSO.
- There one with zero: SO



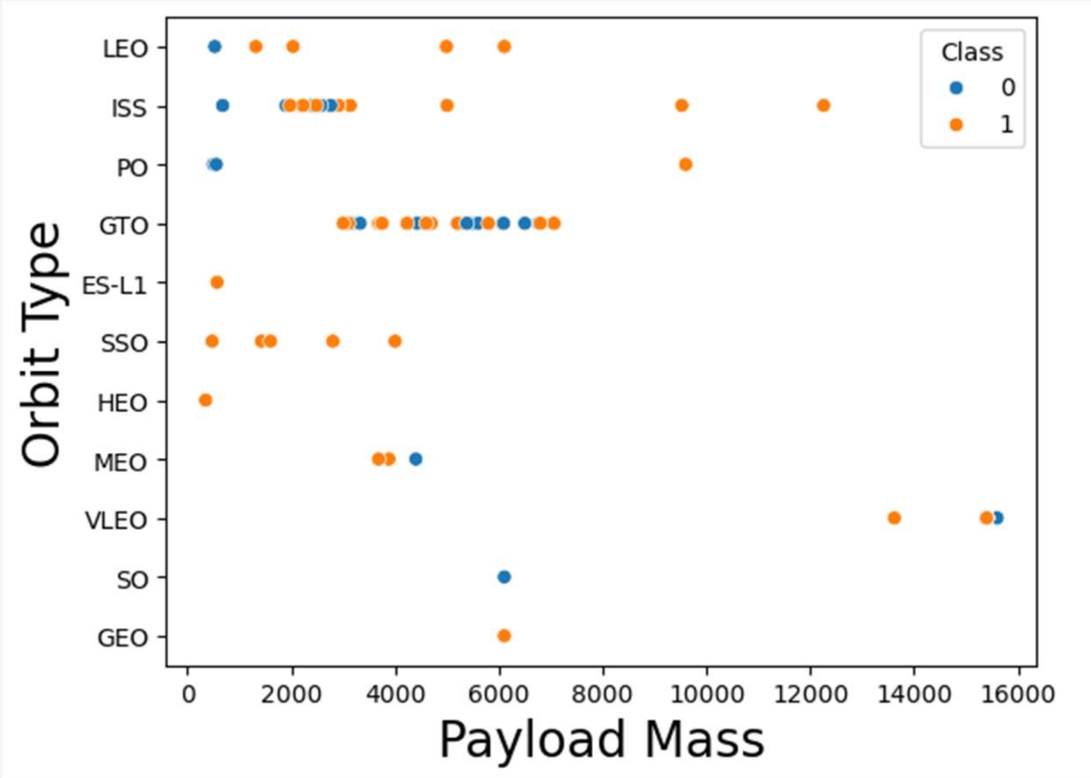
Flight Number vs. Orbit Type

- Here we have a scatter plot for Flight numbers vs Orbit.
- In orange color the success landings and blue the failed landings for each orbit.
- As we can see the zero rate orbit SO only have one landing, as HEO with one success landing, The orbit more used are: ISS,GTO,VLEO



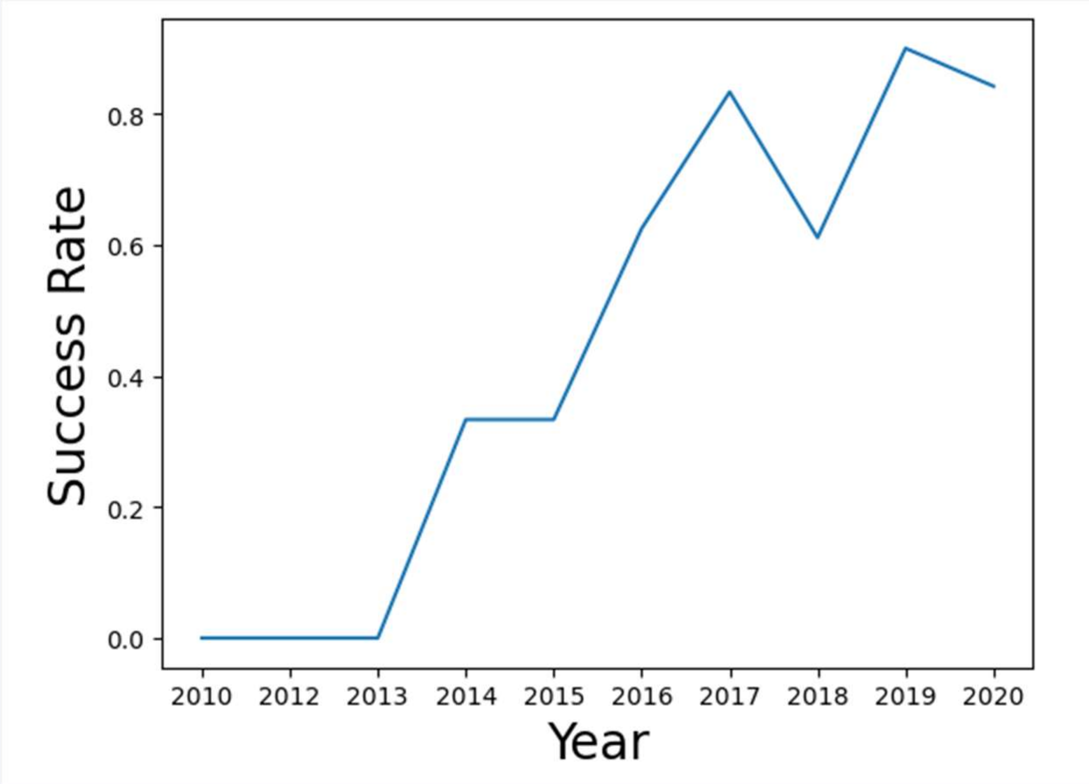
Payload vs. Orbit Type

- Here we have a scatter plot for Payload Mass vs Orbit Type.
- In orange color the success landings and blue the failed landings for each site.



Launch Success Yearly Trend

- Here we have a line plot for Years vs Success Rate.
- We can see the rate became improve in 2013.



All Launch Site Names

- Find the names of the unique launch sites in the space mission:
- SELECT DISTINCT launch_site FROM SPACEXTBL

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

- There are four launch sites in the data base.

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`:

SELECT * FROM SPACEXTBL WHERE launch_site LIKE "CCA%" LIMIT 5

- As we see all records found has a launch site who begin with CCA

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)

Total Payload Mass

- Calculate the total payload carried by boosters from NASA:

SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer LIKE "NASA (CRS)“

SUM(PAYLOAD_MASS__KG_)
45596

- The total payload carried was 45 596.

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1:

```
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE
Booster_Version LIKE "F9 v1.1%“
```

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

- The average payload mass carried by booster version F9 v1.1 was 2534,67

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad:

SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome LIKE "%Success (ground pad)%“

MIN(Date)
2015-12-22

- The first successful landing outcome was 22 / 12 / 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome LIKE "Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- There are four booster version which have successfully landed on drone ship and had payload mass between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes:

```
SELECT Mission_Outcome, COUNT(*) FROM SPACEXTBL GROUP BY
Mission_Outcome
```

Mission_Outcome	COUNT(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- The higher mission outcome was 98.

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass:

```
SELECT Booster_Version FROM SPACEXTBL WHERE
PAYLOAD_MASS__KG_ IN (SELECT MAX(PAYLOAD_MASS__KG_)
FROM SPACEXTBL)
```

- There are twelve booster which have carried the maximum payload mass

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:

SELECT substr(Date, 6,2) AS month, Landing_Outcome , Booster_Version, launch_site FROM SPACEXTBL WHERE substr(Date,0,5)='2015' AND Landing_Outcome LIKE "%Failure (drone ship)%“

month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The month in 2015 with failed landing outcomes in drone ship was January and April

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order:

```
SELECT Landing_Outcome,
COUNT(Landing_Outcome) AS COUNT FROM
SPACEXTBL GROUP BY Landing_Outcome
HAVING Date > DATE("2010-06-04") AND Date
< DATE("2017-03-20") ORDER BY COUNT DESC
```

- In general the success count is greater than failure, the times it has been attempt.

Landing_Outcome	COUNT
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Precluded (drone ship)	1

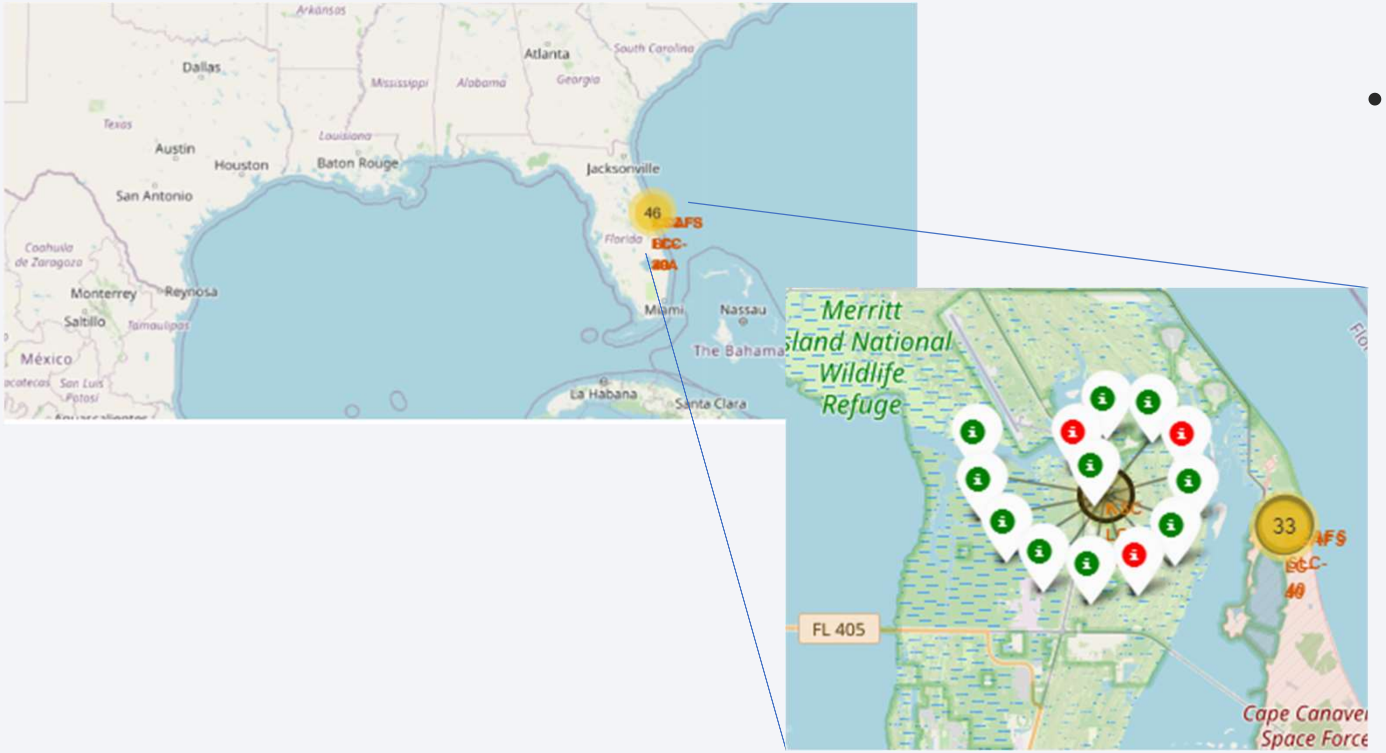
Section 3

Launch Sites Proximities Analysis

Launch Sites Locations

- One site locations is in Texas, the other three in Florida.

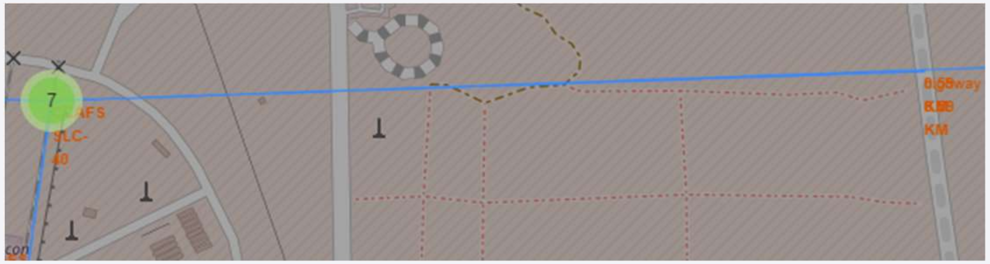
Success/failed landings for each site



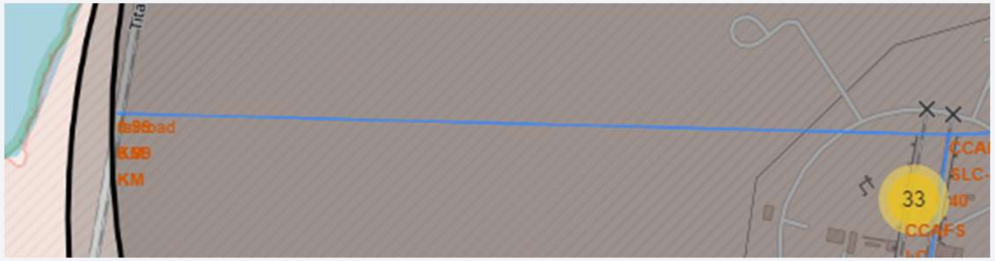
- If we zoom or click in the site we can see the success and failed landings in each site.

Proximities of the launch site: CCAFS SLC 40

- To highway:



- To Railroad:

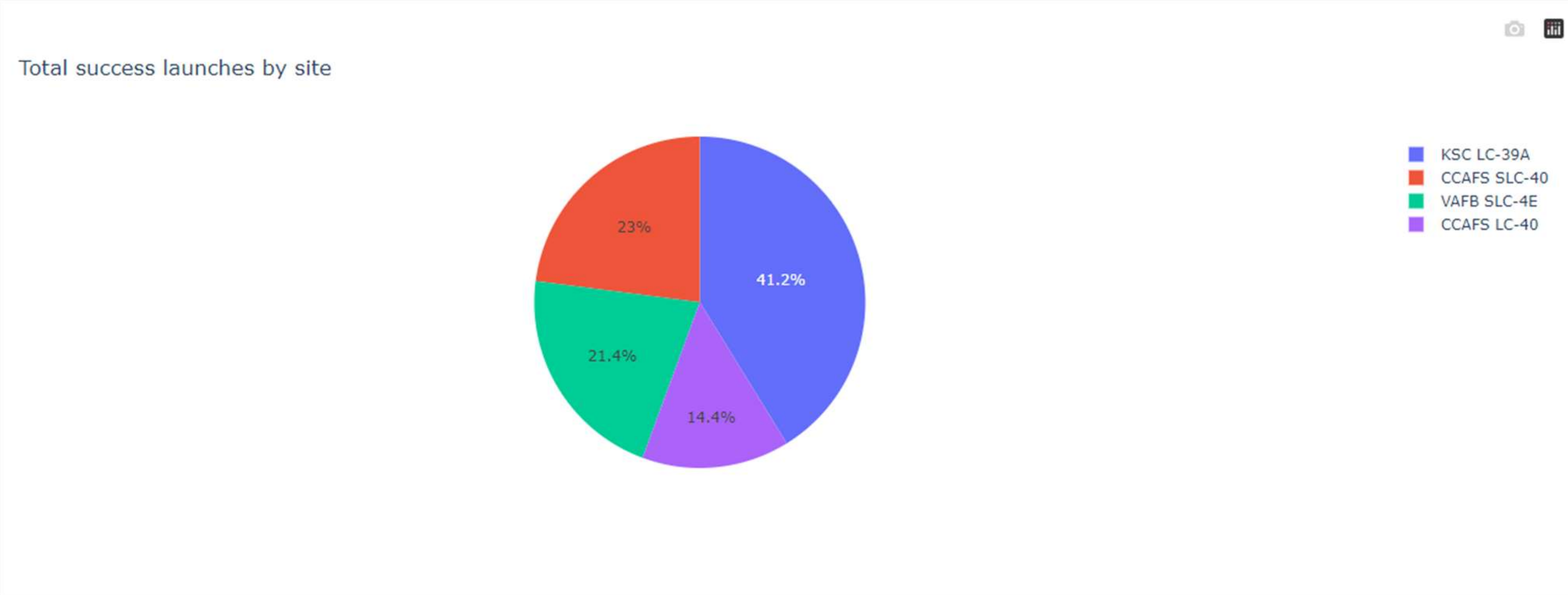


- To city:

Section 4

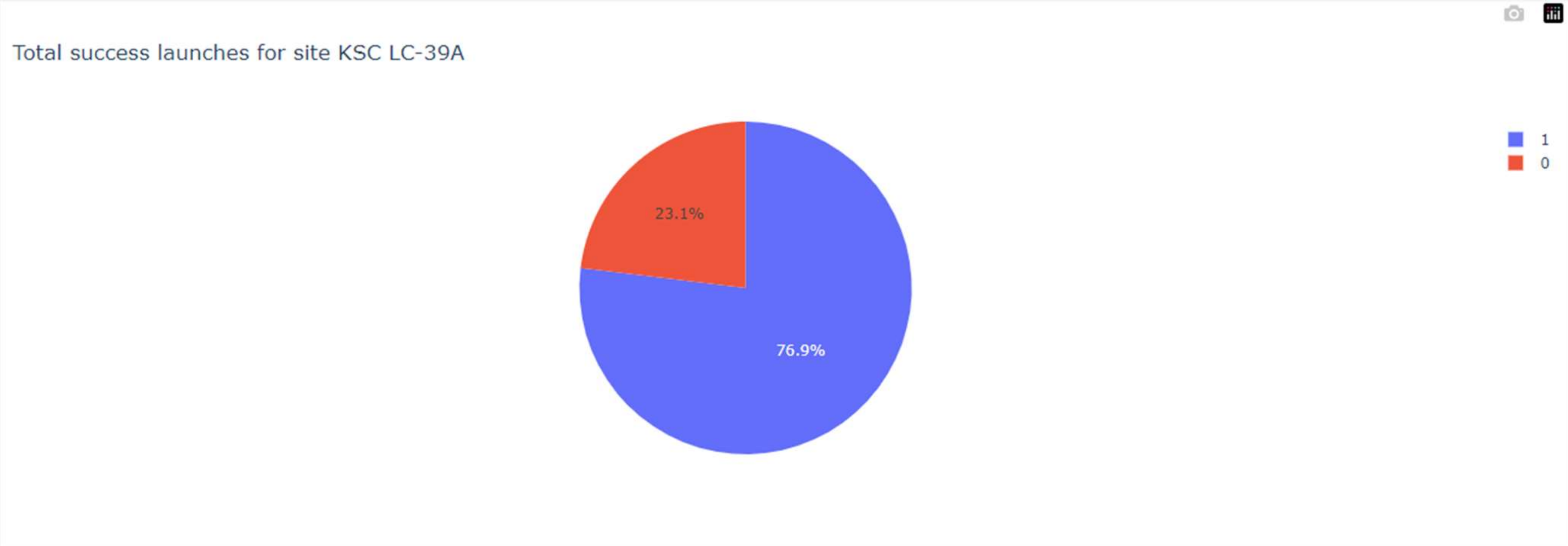
Build a Dashboard with Plotly Dash

Launch success count for all sites



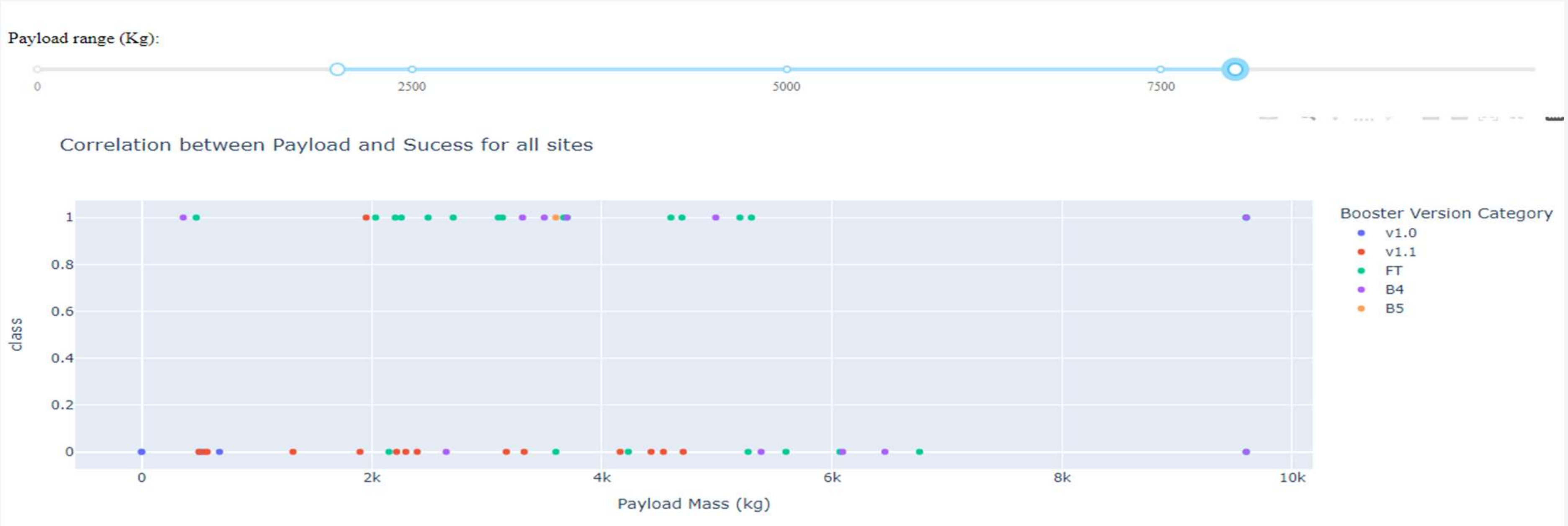
- The site with more success landings is KSC LC-39A and with less success landings: CCAFS LC-40

Total success launches for site KSC LC-39A



- The rate of success of this site is over 76%

Payload vs. Launch Outcome scatter plot for all sites



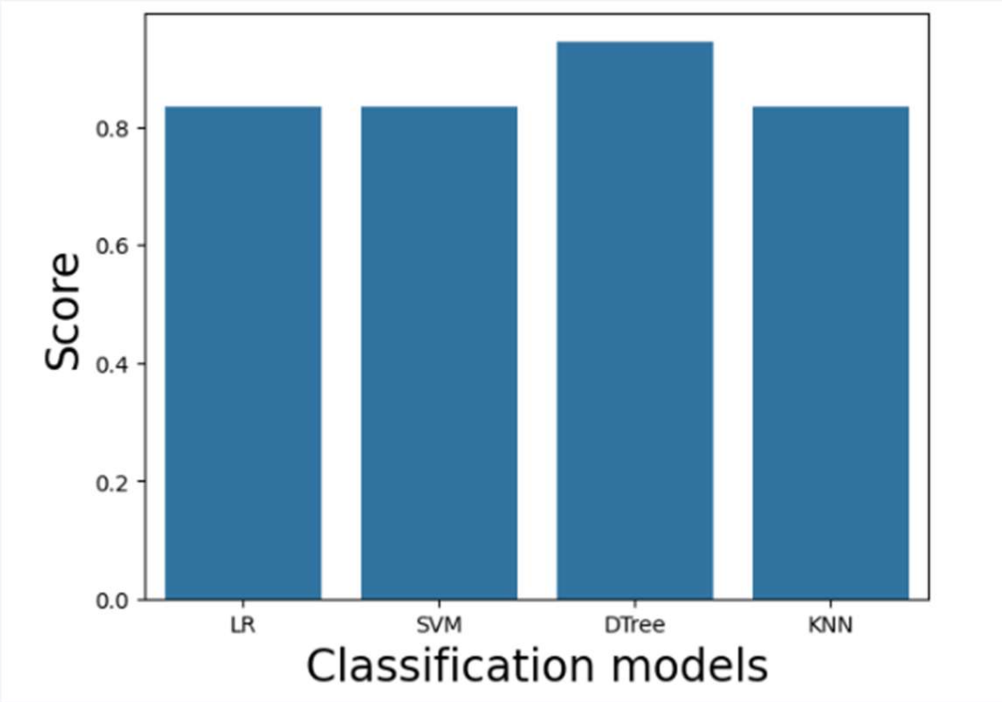
Section 5

Predictive Analysis (Classification)

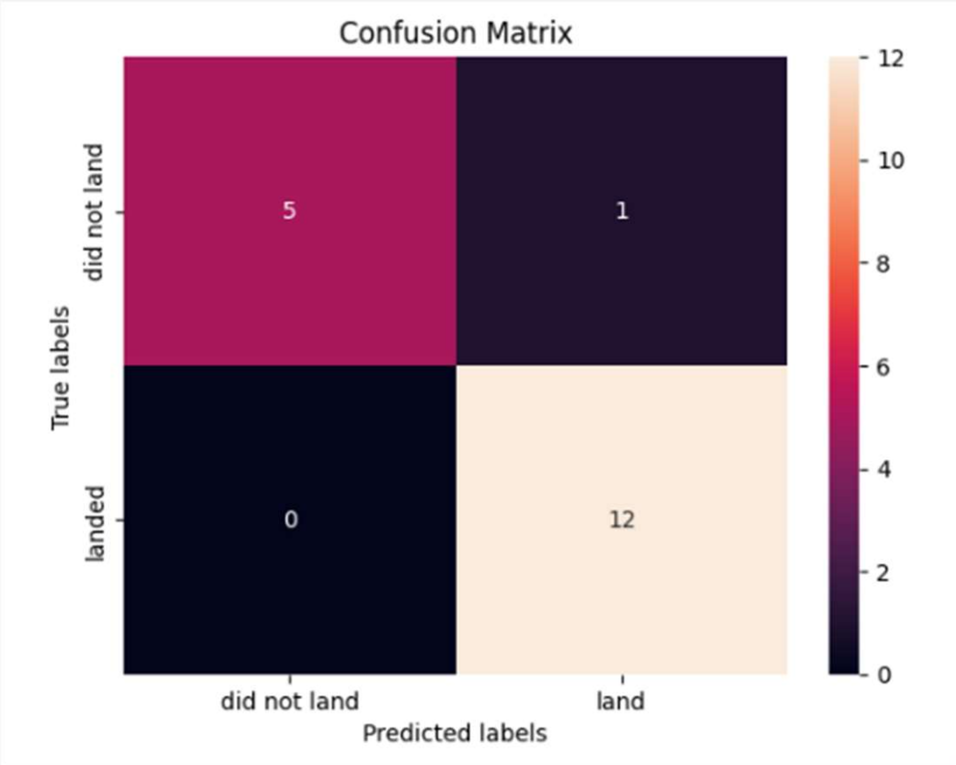
Classification Accuracy

- The model with the highest classification accuracy was the Decision Tree:

0.94



Confusion Matrix



Here we can see for the success was 100% true predicted, and for failed 5 well predicted, only 1 not.

Conclusions

- All the sites are near from highway, railroad, and the coast, but far away from the cities.
- The site with better Success rate is KSC LC 39A
- The Decision Tree Model is the best for predict the landing outcome: Success or Failed with an accuracy of 94%


Thank you!